# Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding

*Year: 2023* | *Citations: 340* | *Authors: Peng Jin, Ryuichi Takanobu, Caiwan Zhang*

## Abstract

Large language models have demonstrated impressive universal capabilities across a wide range of open-ended tasks and have extended their utility to encompass multi-modal conversations. However, existing methods encounter challenges in effectively handling both image and video understanding, particularly with limited visual tokens. In this work, we introduce Chat-UniVi, a Unified Vision-language model capable of comprehending and engaging in conver-sations involving images and videos through a unified visual representation. Specifically, we employ a set of dynamic visual tokens to uniformly represent images and videos. This representation framework empowers the model to ef-ficiently utilize a limited number of visual tokens to simul-taneously capture the spatial details necessary for images and the comprehensive temporal relationship required for videos. Moreover, we leverage a multi-scale representation, enabling the model to perceive both high-level seman-tic concepts and low-level visual details. Notably, Chat-UniVi is trained on a mixed dataset containing both images and videos, allowing direct application to tasks involving both mediums without requiring any modifications. Exten-sive experimental results demonstrate that Chat- UniVi con-sistently outperforms even existing methods exclusively de-signed for either images or videos. Code is available at https://github.com/PKu-Yuan Group/Chat-UniVi.